



Testing multivariate causal hypotheses when you can't do a controlled experiment

When there are no "missing" variables → **D-sep tests**

When there are "missing" variables → **Structural equations with latent variables**

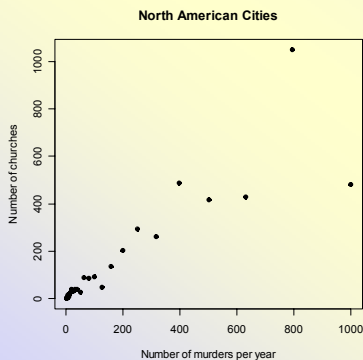
When you want to explore possible hypotheses → **Exploratory path analysis
And
Vanishing tetrad equations**

Département de biologie



Correlation does not imply causation...

But causation DOES imply correlation (almost always)



Population size → **Number of churches**
→ **Number of murders**

How do we experimentally test such causal claims?

Experimental control

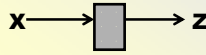
Randomisation

Département de biologie

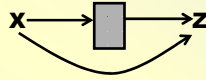




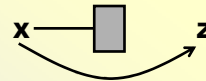
Logic of a controlled experiment



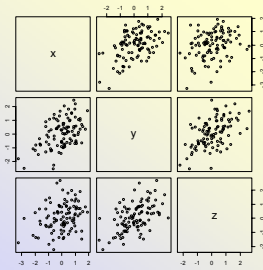
X and Z conditionally Independent



X and Z conditionally Dependent



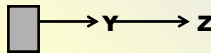
X and Z conditionally independent



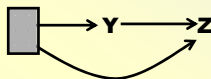
Département de biologie



Logic of a controlled experiment



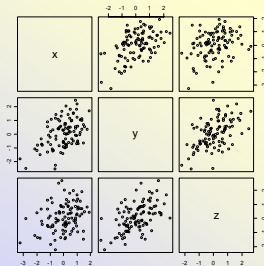
Y and Z conditionally dependent



Y and Z conditionally dependent



Y and Z conditionally independent



Département de biologie





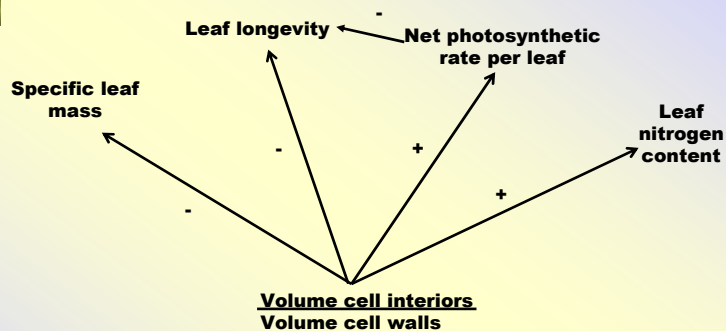
Logic of a controlled experiment

1. **Hypothesise a causal structure between the variables that accounts for observed correlations.**
2. **Determine how this pattern of correlations would change upon controlling different variables by fixing their values to some constant value.**
3. **Physically control these variables, and compare observed and predicted *conditional* correlations given this hypothesised causal process.**

Département de biologie



How can one conduct a controlled experiment to test this Hypothesis in an interspecific (evolutionary) context?



Département de biologie





Logic of the statistical tests I will describe

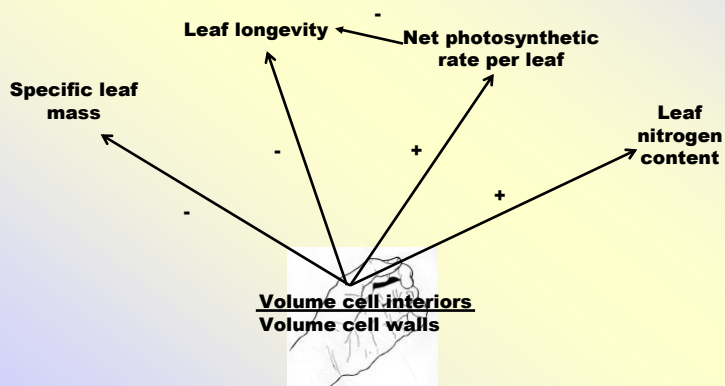
1. **Hypothesise a causal structure between the variables that accounts for observed correlations.**
2. **Determine how this pattern of correlations would change upon controlling different variables by fixing their values to some constant value.**
3. **STATISTICALLY control these variables, and compare observed and predicted *conditional* correlations given this hypothesised causal process.**

Département de biologie



What we need:

Under what conditions will statistical control give the same Answer as experimental control?



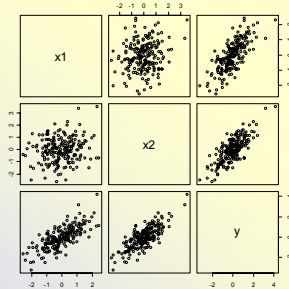
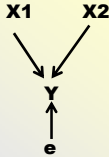
Département de biologie



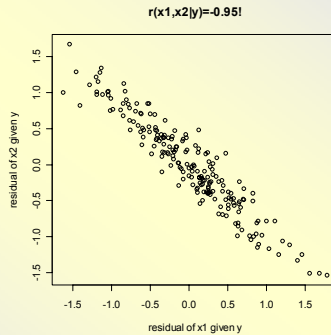


Problem

Statistical control does not always result in the same pattern of conditional correlations as physical control



Is X1 independent of X2, statistically holding constant Y?



Département de biologie



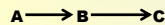
Pearl, J. 2000. *Causality. Models, Reasoning, and Inference*. Cambridge U. Press, Cambridge.



Spirtes, P. et al. 1993. *Causation, Prediction, and Search*. Springer-Verlag, NY.



1. Express causal claims using graph theory (directed acyclic graphs - DAGs)
Property: asymmetric relationships



2. Apply a graph-theoretic operator (d-separation) on this graph.

$A \perp\!\!\!\perp C | B$ (A is separated from C given B in the graph)

3. If two vertices (X, Y) in this DAG are d-separated given a set Q of other vertices, then variables X & Y will be probabilistically independent conditional on Q in EVERY multivariate probability distribution that is generated by the causal process represented by the DAG.

Département de biologie





D-separation

A directed path p between X and Y is **d-separated** (or **blocked**) by a set of other **conditioning variables** $Z=\{A,B,\dots\}$ if and only if

$$X \longrightarrow \dots \longleftarrow M \longrightarrow \dots \longrightarrow Y$$

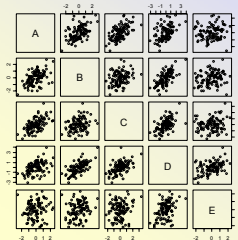
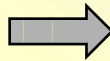
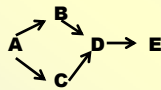
1. p contains a **chain** ($i \rightarrow m \rightarrow j$) or a **fork** ($i \leftarrow m \rightarrow j$) such that the middle variable (m) is in the conditioning set Z , or
2. p contains a **collider** ($i \rightarrow m \leftarrow j$) such that the middle variable (m) is **not** in the conditioning set Z and such that no causal descendent (effect) of m is in the conditioning set Z

Two variables (X, Y) in a directed acyclic graph are **d-separated** given the Conditioning set Z if and only if Z blocks every path from variable X to variable Y .

Département de biologie



IF a set of observed data are generated according to a causal process described by a directed acyclic graph



THEN, no matter what the probability distribution of the variables, or the functional form of the links between the variables,

D-separation in the graph implies (conditional) independence in the data, and *vice versa*.

So, if two variables are not (conditionally independent in your data, but your causal hypothesis – in the form of a directed graph – predicts d-separation, then your causal hypothesis is wrong!

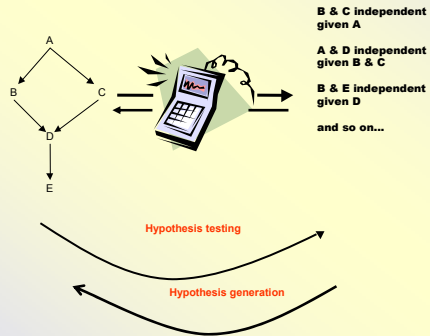
Département de biologie





"3-D" causal process

"2-D" correlational shadow



Département de biologie



D-sep tests

Advantages:

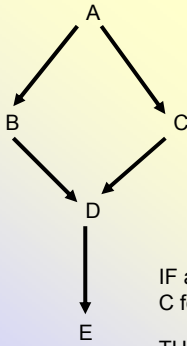
- Can be adapted to any multivariate probability distribution
- Can be adapted to any functional (acyclic) relationship between variables.

Disadvantages:

Your causal hypothesis cannot have unmeasured (« latent ») variables (with a few exceptions...)

Département de biologie





List basis set B_U

- $A \perp\!\!\!\perp D \mid \{B, C\}$
- $A \perp\!\!\!\perp E \mid \{D\}$
- $B \perp\!\!\!\perp C \mid \{A\}$
- $B \perp\!\!\!\perp E \mid \{A, D\}$
- $C \perp\!\!\!\perp E \mid \{A, D\}$

Convert to probabilistic claims

- $r_{A,D \mid \{B,C\}} = 0$
- $r_{A,E \mid D} = 0$
- $r_{B,C \mid A} = 0$
- $r_{B,E \mid \{A,D\}} = 0$
- $r_{C,E \mid \{A,D\}} = 0$

Calculate probability of each claim in data

- $p_1 = 0.23$
- $p_2 = 0.50$
- $p_3 = 0.001$
- $p_4 = 0.45$
- $p_5 = 0.12$

Calculate: $C = -2 \sum_{i=1}^k \text{Ln}(p_i)$

$C = 23.98$
 $k = 5$

IF all d-sep claims in the graph are true in the data, then C follows a chi-squared distribution with $2k$ degrees of freedom

THEREFORE if the probability of C is **below the significance level**..... the **causal structure is rejected** by the data.

THEREFORE if the probability of C is **above the significance level**..... the **causal structure is consistent** with the data.

Département de biologie 23.98 with 10 degrees of freedom gives $p=0.008$



REJECT causal structure



Modelling interacting traits



Prunus mahaleb

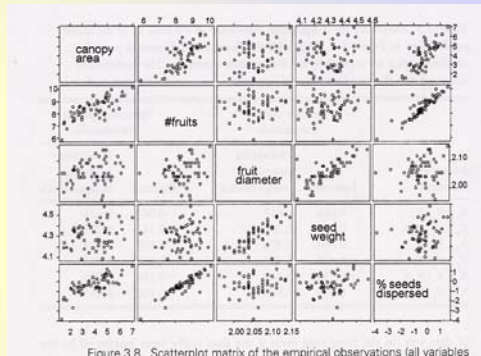


Figure 3.8 Scatterplot matrix of the empirical observations fall variables

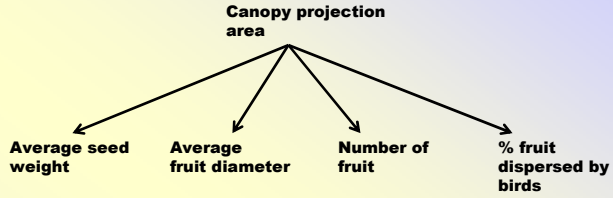
Jordano, P. (1995). **Fruivore-mediated selection on fruit and seed size: birds and St. Lucie's Cherry.** *Ecology* 76:2627-2639

Département de biologie

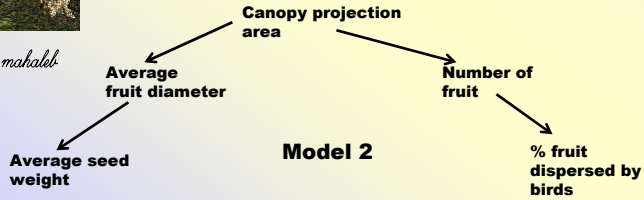




Prunus mahaleb



Model 1

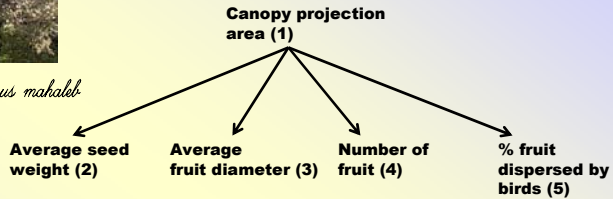


Model 2

Département de biologie



Prunus mahaleb



Model 1

Basis set	Predicted independencies	Values (probabilities)
2 3 1	$r_{2,3 1}=0$	0.02 (0.87)
2 4 1	$r_{2,4 1}=0$	0.11 (0.42)
2 5 1	$r_{2,5 1}=0$	0.80 ($<10^{-4}$)
3 4 1	$r_{3,4 1}=0$	0.77 ($<10^{-4}$)
3 5 1	$r_{3,5 1}=0$	-0.08(0.57)
4 5 1	$r_{4,5 1}=0$	0.02 (0.87)

$C=137.64, 12 \text{ df}, p<10^{-5}$

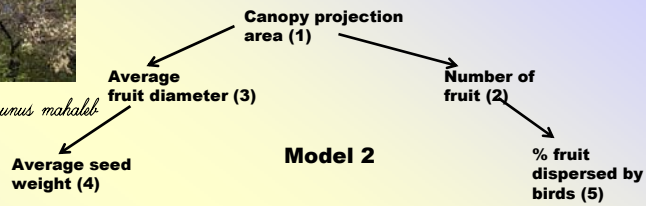
Département de biologie

Data falsify the model





Prunus mahaleb



Model 2

Basis set	Predicted independencies	Values (probabilities)
1 4 3	$r_{1,4 3}=0$	0.07 (0.62)
1 5 2	$r_{1,5 2}=0$	0.08 (0.57)
2 3 1	$r_{2,3 1}=0$	0.02 (0.87)
3 4 (1,3)	$r_{2,4 (1,3)}=0$	0.14 (0.29)
3 5 (1,2)	$r_{3,5 (1,2)}=0$	-0.16 (0.25)
4 5 (2,3)	$r_{4,5 (2,3)}=0$	0.01 (0.97)

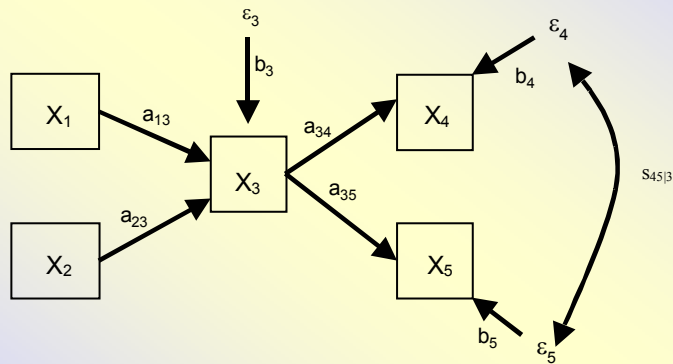
C=7.69, 12 df, p=0.81

Département de biologie Data consistent with this model



Testing causal models with unmeasured (latent) variables - Structural Equation Models (SEM)

STEP 1: write down the causal model as a directed graph



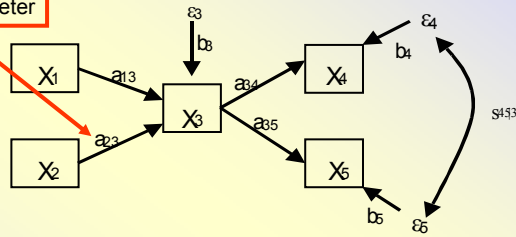
OPTIONAL STEP: center each variable about its mean ($X_i - \bar{X}_i$)

Département de biologie





free parameter



STEP 2: translate this to a series of structured equations with free parameters

$$X_1 = N(0, \sigma) \quad \varepsilon_3 = N(0, 1) \quad \varepsilon_5 = N(0, 1)$$

$$X_2 = N(0, \sigma) \quad \varepsilon_4 = N(0, 1)$$

$$X_3 = a_{13}X_1 + a_{23}X_2 + b_3\varepsilon_3 \quad X_4 = a_{34}X_3 + b_4\varepsilon_4 \quad X_5 = a_{35}X_3 + b_5\varepsilon_5$$

$$\text{Cov}(X_1, X_2) = \text{Cov}(X_1, \varepsilon_3) = \text{Cov}(X_1, \varepsilon_4) = \text{Cov}(X_1, \varepsilon_5) = \text{Cov}(X_2, \varepsilon_3) = \text{Cov}(X_2, \varepsilon_4) = \text{Cov}(X_2, \varepsilon_5) =$$

$$\text{Cov}(\varepsilon_3, \varepsilon_4) = \text{Cov}(\varepsilon_3, \varepsilon_5) = 0$$

$$\text{Cov}(\varepsilon_4, \varepsilon_5) = \sigma_{45}$$

Département de biologie



STEP 3: Derive the predicted variance and covariance between each pair of variables in the model, respecting the constraints implied by the causal graph, using covariance algebra.

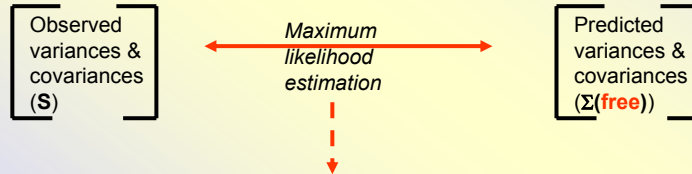
$$\Sigma = \begin{matrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{matrix} \begin{bmatrix} S_1 & & & & \\ & \mathbf{0} & & & \\ & & S_2 & & \\ & & & S_3 & \\ & & & & S_4 \\ & & & & & S_5 + b_4 b_5 \text{Cov}(\varepsilon_4, \varepsilon_5) \end{bmatrix}$$

Département de biologie





STEP 4: Estimate the free parameters by **minimizing** the difference between the observed and predicted variances and covariances.



Estimates of the free parameters (path coefficients, error variances, free covariances) that make the predicted covariance matrix (Σ) as **close as possible** to the observed variances and covariances, while **respecting the constraints** required by d-separation (i.e. the causal structure).

Département de biologie



STEP 6: Look at the remaining differences between the observed and predicted covariance matrices, and calculate the probability of having observed this difference, assuming that these differences should be the same except for random sampling variation.

X^2_{ML} (maximum likelihood chi-square statistic)

degrees of freedom = $V(V+1)/2$ - #free parameters

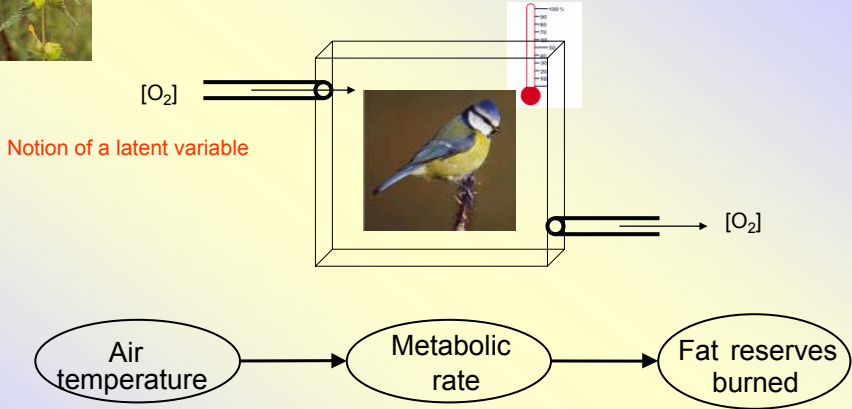
STEP 7: If the calculated probability is less than the chosen significance level (eg 0.05) then the data did not come from this causal process, and the model must be rejected; otherwise the data support the model.

Département de biologie

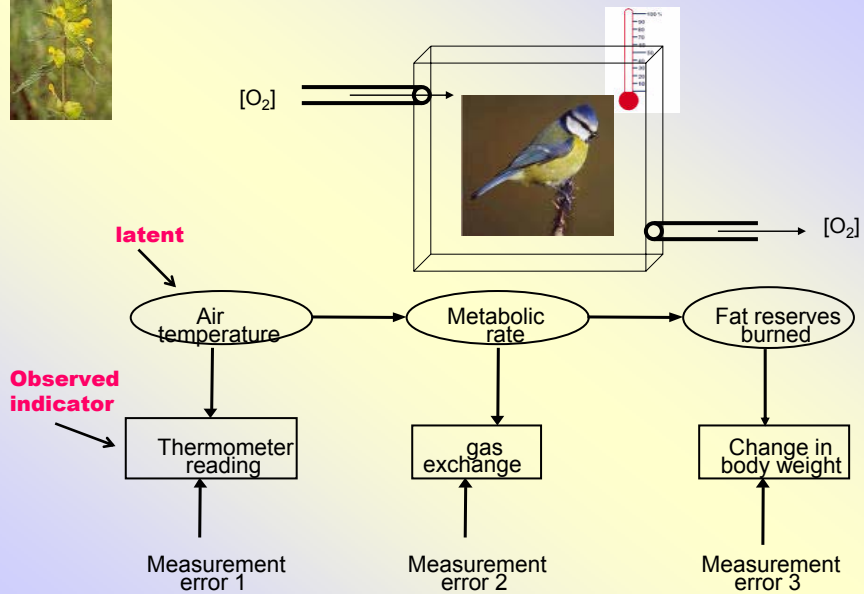




Measurement models and maximum likelihood

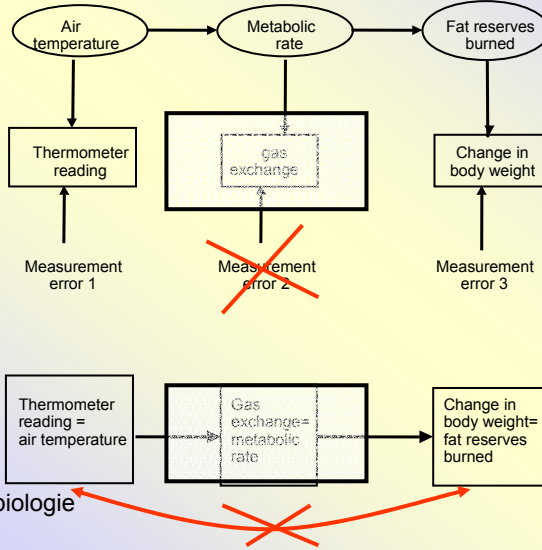


Département de biologie



Département de biologie





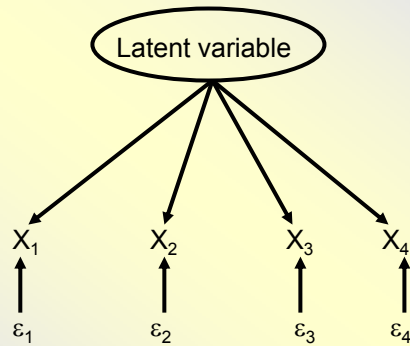
Département de biologie



Latent variable

Observed (indicator) variables

Error variables

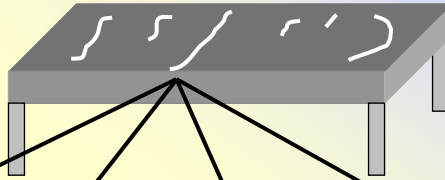


Département de biologie





True length
of strings
(latent)



Ruler
 $\pm 1\text{ cm}$



Her hand
 $\pm 0.07\text{ hand}$

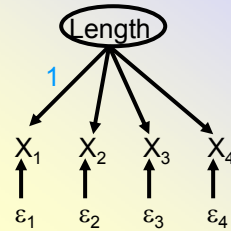


Ruler
 $\pm 1\text{ inch}$



Visual estimation
 $\pm 10\text{ cm}$

Département de biologie



$$L = N(0, \sigma) \quad \varepsilon_1 = N(0, \sigma) \quad \varepsilon_2 = N(0, \sigma) \quad \varepsilon_3 = N(0, \sigma) \quad \varepsilon_4 = N(0, \sigma)$$

$$X_1 = 1L + \varepsilon_1 \quad \text{Cov}(\varepsilon_1, \varepsilon_2) = \text{Cov}(\varepsilon_1, \varepsilon_3) = \text{Cov}(\varepsilon_1, \varepsilon_4) = \text{Cov}(\varepsilon_2, \varepsilon_3) =$$

$$X_2 = a_2L + \varepsilon_2 \quad \text{Cov}(\varepsilon_2, \varepsilon_4) = \text{Cov}(\varepsilon_3, \varepsilon_4) = 0$$

$$X_3 = a_3L + \varepsilon_3$$

$$X_4 = a_4L + \varepsilon_4$$

Département de biologie





**Wright, I.J. 2004. The worldwide leaf economics spectrum.
Nature 428:821-827.**

**2548 species from 219 families
representing 175 sites from the arctic to the tropics**

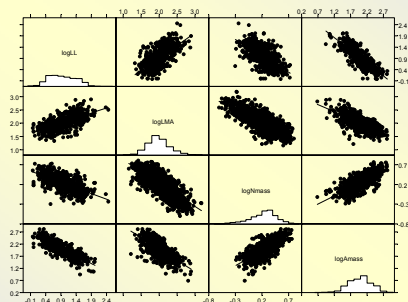
**maximum photosynthetic rate (A_{max})
leaf mass per area (LMA)
nitrogen content per mass (N_{mass})
leaf longevity (LL)**

Given the incredible variation in leaf form and physiological performance, given the deep phylogenetic origin of this variation, and given the wide variety of environments in which leaves operate, one might expect many different types of tradeoffs involving leaf form and function. If so, then the patterns of correlation among functional leaf traits should vary greatly across differing phylogenies and environments.

Département de biologie



**Wright, I.J. 2004. The worldwide leaf economics spectrum.
Nature 428:821-827.**



**82% of the total variance in these 4 variables is
Accounted for by a single principal component.**

**Patterns of correlation are largely independent of
Phylogenetic details or habitat differences!**

Département de biologie

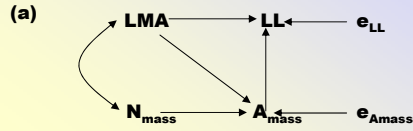




**Two alternative hypotheses that have been suggested
To account for this pattern of covariation**

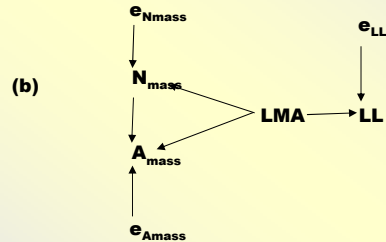
Wright et al. 2004

$X^2=12.83, 1df, p=0.003$



Meziane & Shipley 2001

$X^2=209.37, 2 df, p<10^{-15}$



Département de biologie



**Algorithm to exhaustively explore all possible orderings
Of variables, assuming no latent variables**

Pearl, J. 2000. *Causality. Models, Reasoning, and Inference.* Cambridge U. Press, Cambridge.



CI (causal inference) algorithm

Spirtes, P. et al. 1993. *Causation, Prediction, and Search.* Springer-Verlag, NY.



FCI algorithm

Test each possible one using d-sep test: Result

There is no such model that passes the test!

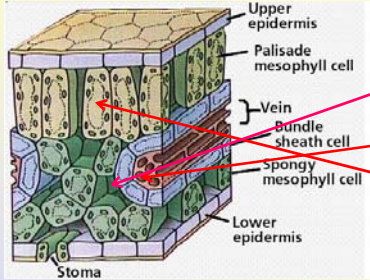
Département de biologie





An alternative hypothesis that involves:

- (i) unmeasured variation in cell size and cell wall thickness**
- (ii) selection on leaf longevity to maximise plant carbon gain rather than leaf-level carbon gain.**

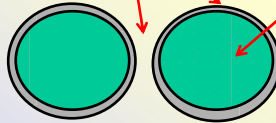


Leaf volume =

Volume of air spaces

Volume occupied by cell walls

Volume bounded by cell membranes

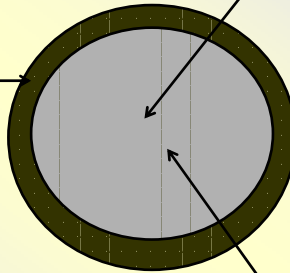


Département de biologie



Most of the cell water is inside the cell membrane

Most of the dry mass is in the cell wall; i.e. proportional to cell surface area



Most of the nitrogen and metabolic activity (i.e. photosynthesis) is inside the cell membrane; i.e. Proportional to cell volume & water mass

$$\frac{V_{\text{cell}}}{V_{\text{wall}}} \rightarrow \frac{\text{Mass}_{\text{water}}}{\text{Dry mass}}$$

Département de biologie





Leaf nitrogen per mass (N_{mass}):

$$N_{\text{mass}} = \frac{N}{(M_{\text{cell}} + M_{\text{wall}})} \cong \frac{\bar{n}V_{\text{cell}}}{(M_{\text{cell}} + M_{\text{wall}})}$$
 . If the cell does not contain large quantities of non-structural carbohydrates then $M_{\text{wall}} \gg M_{\text{cell}}$ and

$$N_{\text{mass}} \cong \frac{\bar{n}V_{\text{cell}}}{dV_{\text{wall}}}$$
 where d (specific gravity) is relatively constant among species at 1.5

$$\frac{V_{\text{cell}}}{V_{\text{wall}}} \rightarrow N_{\text{mass}}$$

Leaf net photosynthesis per mass (A_{mass}):

\bar{a} is the average rate of net carbon fixation per cell volume, then

A (total net carbon fixation per leaf) = $\bar{a}V_{\text{cell}}$. Therefore,

A_{mass} (total net carbon fixation per leaf mass) = $A_{\text{mass}} = \frac{A}{M_{\text{cell}} + M_{\text{wall}}} \cong \frac{\bar{a}V_{\text{cell}}}{dV_{\text{wall}}}$

Département de biologie

$$\frac{V_{\text{cell}}}{V_{\text{wall}}} \rightarrow A_{\text{mass}}$$



LMA (leaf mass per area)

$LMA = 1/SLA$; ($SLA = \text{leaf area per mass}$)

$$SLA = \frac{S}{M_{\text{cell}} + M_{\text{wall}}} = \frac{V_L}{T(M_{\text{cell}} + M_{\text{wall}})} \cong \frac{V_L}{T(dV_{\text{wall}})}$$

$$SLA \cong \frac{V_{\text{cell}} + V_{\text{wall}} + V_{\text{air}}}{T(dV_{\text{wall}})} = \frac{1}{d} \left(\frac{V_{\text{cell}}}{V_{\text{wall}}} + \frac{1}{T} + \frac{V_{\text{air}}}{V_{\text{wall}}} \right)$$

$$\frac{V_{\text{cell}}}{V_{\text{wall}}} \rightarrow SLA = 1/LMA$$



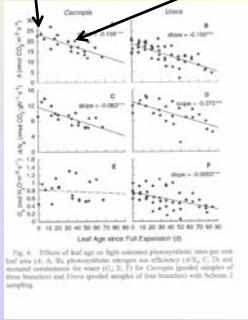
Leaf lifespan (LL)

Kikuzawa's (1991) model:

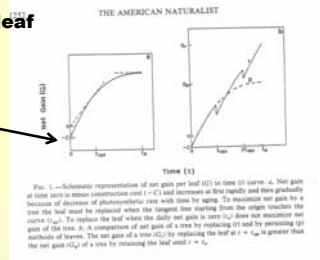
$$LL = \sqrt{\frac{2bC}{A_{max}}}$$

A_{max} = maximum net photosynthetic rate when
The leaf first becomes autotrophic

b = rate of decrease in A as the leaf ages



C = construction cost of the leaf
(amount of glucose used to construct the leaf)
~ carbon content



Département de biologie



$$C_M = \frac{\Delta}{(M_{cell} + M_{wall})} \cong \frac{\Delta}{dV_{wall}} \cong \sum_i \delta_{0i} \bar{e}_i \frac{V_{cell}}{dV_{wall}} + \frac{1}{d} \sum_j \delta_{0j} \bar{e}_j$$

The construction cost needed to absorb, fix, transport or biochemically manipulate each atom of element i in the liquid phase

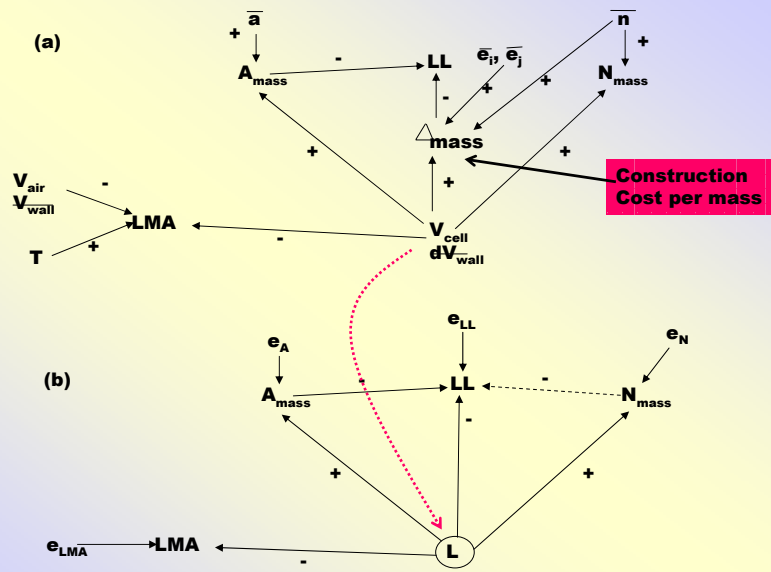
Same thing for elements in solid phase (cell walls)

Concentration per volume of liquid phase elements

$$\frac{V_{cell}}{V_{wall}} \rightarrow CC$$

Département de biologie

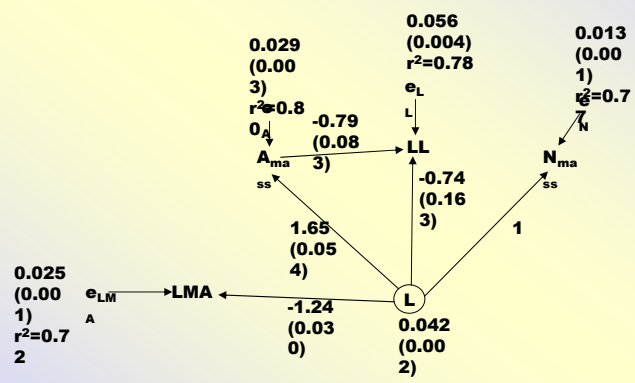




Département de biologie



$\chi^2=0.766$,
1 df,
P=0.38



Département de biologie





For more information:



Shipley, B. 2000. Cause and correlation in Biology: A user's guide to path analysis, Structural equations and causal inference. Cambridge University Press.

Département de biologie

